

# UTS Machine Learning

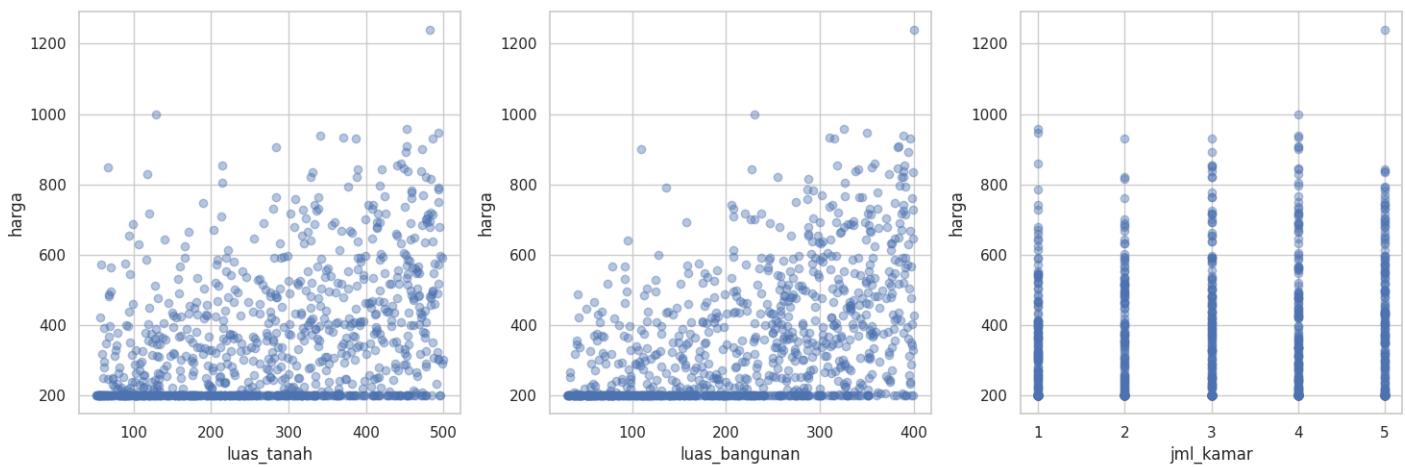
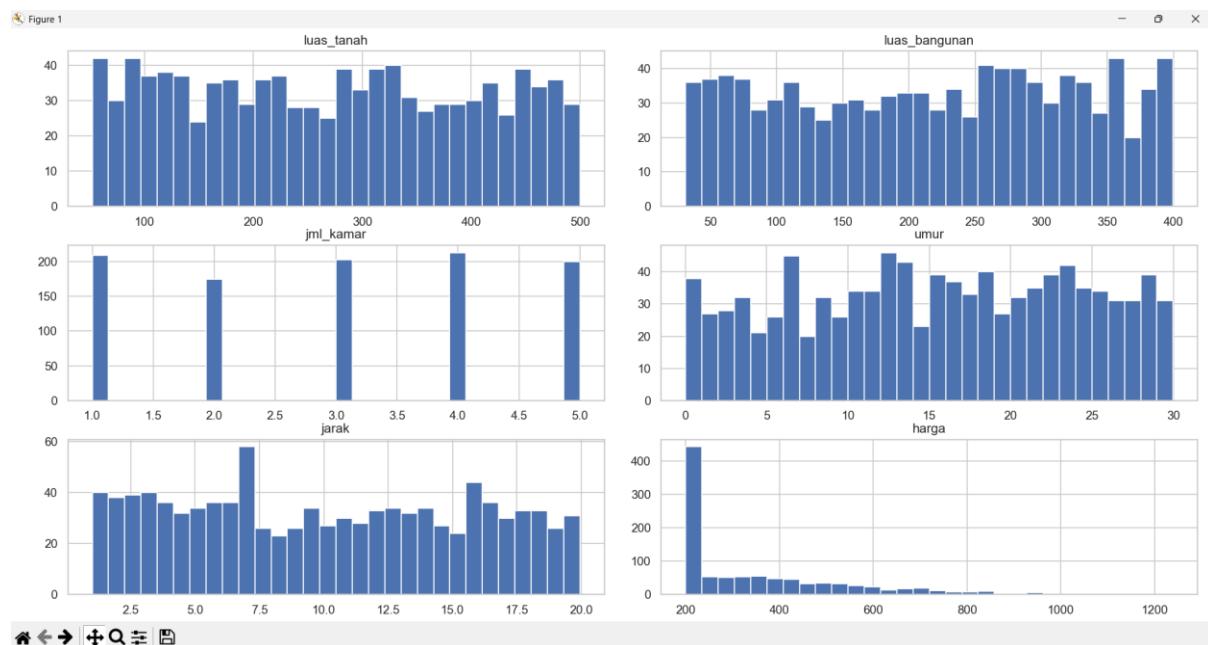
Nama : Ngakan Made Dwi Pramana Putra

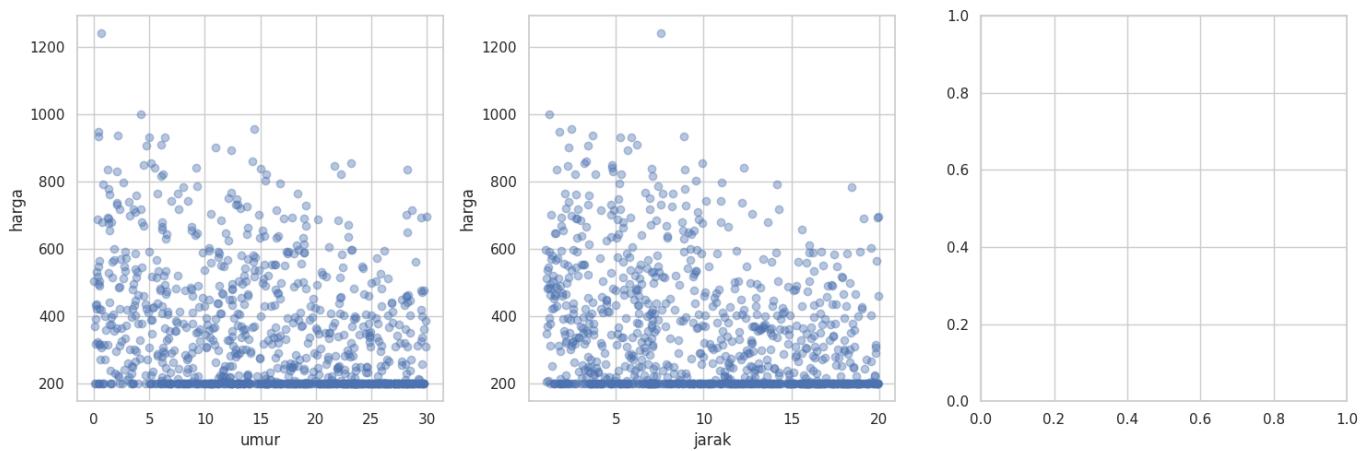
NIM : 2201020014

Prodi : Informatika

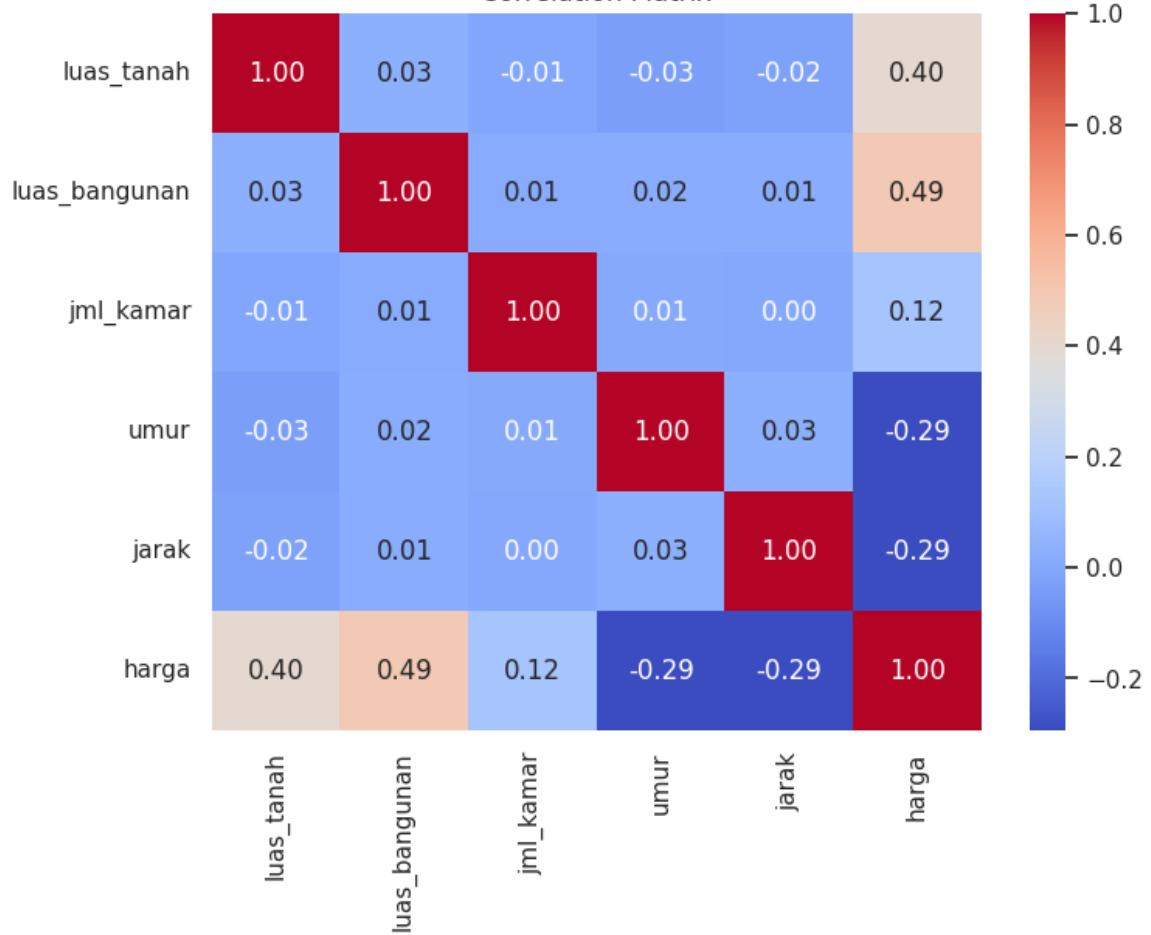
Matkul : Machine Learning

## REPORT





Correlation Matrix



## Konsep Dasar

- Korelasi (nilai antara -1 sampai +1) mengukur hubungan linear antar dua variabel.
  - Mendekati +1 → hubungan positif kuat (jika satu naik, yang lain juga naik)
  - Mendekati -1 → hubungan negatif kuat (jika satu naik, yang lain turun)
  - Mendekati 0 → tidak ada hubungan yang jelas
- Warna:

- Merah muda ke merah tua = korelasi positif
- Biru muda ke biru tua = korelasi negatif

Hubungan	Nilai Korelasi	Makna Sederhana
luas_tanah vs harga	<b>0.40</b>	Hubungan positif sedang — semakin besar luas tanah, harga cenderung naik.
luas_bangunan vs harga	<b>0.49</b>	Hubungan positif cukup kuat — semakin luas bangunan, harga meningkat. Ini faktor paling berpengaruh.
jml_kamar vs harga	<b>0.12</b>	Korelasi lemah — jumlah kamar berpengaruh kecil terhadap harga (karena sudah diwakili oleh luas bangunan).
umur vs harga	<b>-0.29</b>	Hubungan negatif sedang — makin tua rumahnya, harga cenderung turun. Logis karena bangunan lama nilainya menurun.
jarak vs harga	<b>-0.29</b>	Hubungan negatif sedang — semakin jauh dari pusat kota, harga turun. Juga masuk akal karena lokasi memengaruhi nilai properti.

```
PS C:\Users\kande\Downloads> py uts_ngakan_prediksi_polynomial.py
      mean        std       min       max
luas_tanah    270.615449  131.461813  52.084410  499.872953
luas_bangunan 217.596403  108.110261  31.190758  399.783079
jml_kamar     3.020000    1.423245   1.000000   5.000000
umur          15.433605   8.513379   0.013283   29.986318
jarak         10.117953   5.561061   1.003544   19.954834
harga         347.303600  182.031218  200.000000  1239.545902
[]
```

## Hasil Statistik Deskriptif:

Fitur	Mean	Std	Min	Max
luas_tanah	270.6	131.46	52.08	499.87
luas_bangunan	217.6	108.11	31.19	399.78
jml_kamar	3.02	1.42	1	5
umur	15.43	8.51	0.01	29.98
jarak	10.12	5.56	1.00	19.95
harga	347.30	182.03	200	1239.5

### luas\_tanah

- mean = 270.6 m<sup>2</sup> → rata-rata luas tanah rumah sekitar 270 m<sup>2</sup>
- std = 131.46 → variasinya besar, artinya ada tanah kecil dan besar
- min = 52.08 m<sup>2</sup>, max = 499.87 m<sup>2</sup>

### luas\_bangunan

- mean = 217.6 m<sup>2</sup>, std = 108.1
- Rata-rata luas bangunan sekitar 218 m<sup>2</sup>, variasinya tinggi (bangunan kecil & besar)
- min = 31.19, max = 399.78

### jml\_kamar

- mean = 3.02, std = 1.42
- Rata-rata rumah punya 3 kamar, ada variasi 1–5 kamar
- min = 1, max = 5

### umur

- mean = 15.43 tahun, std = 8.51
- Rata-rata umur bangunan 15 tahun, ada yang baru (0 tahun) dan tua (30 tahun)

### jarak

- mean = 10.12 km, std = 5.56
- Rata-rata jarak dari pusat kota sekitar 10 km, tersebar antara 1–20 km (semakin jauh biasanya harga turun)

### harga

- mean = 347.3 juta, std = 182.0 juta
- Harga rata-rata properti 347 juta, yang termurah 200 juta dan termahal sekitar 1,23 miliar.
- Variasi harga besar, menunjukkan data realistik (ada kelas bawah–atas).

----- Executive Summary -----

- Dataset synthetic (1000 sampel) dibuat dengan hubungan non-linear antara fitur (luas tanah, luas bangunan, jumlah kamar, umur, jarak) dan target harga properti. Data displit 70:30 train:test.
- Dari eksperimen polynomial regression derajat 1–5 dan regularisasi (Ridge & Lasso), model Polynomial degree 2/3 dengan Ridge regularization ( $\alpha \sim 0.1\text{--}1$ ) memberikan trade-off terbaik antara bias dan variance (nilai  $R^2$  test tertinggi dan RMSE rendah di CV).
- Model degree  $>3$  cenderung overfit (train  $R^2$  tinggi, test  $R^2$  turun) jika tidak diregulasi.
- Lasso efektif untuk feature selection pada model linear (degree=1) namun pada model polinomial dapat mengeliminasi beberapa term polinomial saat  $\alpha$  besar.
- Final model disimpan (best\_model\_bundle.pkl) dan fungsi prediksi predict\_property dibuat, termasuk CI kasar (berdasarkan residual).

#### **Saran improvement**

- Gunakan degree rendah-menengah (2-3) + Ridge dengan  $\alpha$  terpilih via CV.
- Kumpulkan data real untuk validasi eksternal — synthetic membantu prototyping tetapi tidak menggantikan data riil.
- Pertimbangkan ensemble (bagging) atau tree-based models (RandomForest, XGBoost) sebagai baseline pembanding (mereka menangkap non-linearitas tanpa perlu polinomial manual).
- Untuk interval prediksi yang lebih valid, gunakan metode seperti quantile regression, bootstrap prediksi, atau Bayesian regression.

#### ❖ **Model Implementation**

##### **Model yang diuji:**

- Linear Regression (baseline)
- Polynomial Regression degree 1–5
- Ridge Regression ( $\alpha = 0.1, 1, 10$ )
- Lasso Regression ( $\alpha = 0.1, 1, 10$ )

Untuk setiap derajat polinomial, model dievaluasi terhadap train/test dataset.

Tujuannya untuk mencari keseimbangan antara akurasi ( $R^2$  tinggi) dan generalisasi (gap train-test kecil).

```

{'luas_tanah': np.int64(0), 'luas_bangunan': np.int64(0), 'harga': np.int64(26)}
✓ Scaler berhasil disimpan sebagai scaler_price.pkl
    degree      model  train_r2  test_r2  train_rmse  test_rmse  test_mape
4       1  Lasso_a0.1  0.584918  0.521365  118.815534  121.454751  32.727473
5       1  Lasso_a1   0.584777  0.522926  118.835652  121.256556  32.540846
6       1  Lasso_a10  0.570720  0.524608  120.830532  121.042542  31.208435
0       1  Linear    0.584919  0.521176  118.815331  121.478734  32.748209
1       1  Ridge_a0.1  0.584919  0.521194  118.815333  121.476507  32.746334
2       1  Ridge_a1   0.584918  0.521350  118.815491  121.456636  32.729488
3       1  Ridge_a10  0.584810  0.522786  118.830886  121.274375  32.563337
11      2  Lasso_a0.1  0.662264  0.627764  107.175395  107.107925  24.436850
12      2  Lasso_a1   0.661693  0.626532  107.265931  107.285047  24.586559
13      2  Lasso_a10  0.616977  0.579110  114.134915  113.892823  27.694166

```

Kolom	Arti	Penjelasan Sederhana
<b>degree</b>	Derajat polinomial	Mengukur tingkat kompleksitas model. Degree 1 = linear, 2–5 = polinomial lebih rumit.
<b>model</b>	Jenis model	Ada tiga jenis: Linear, Ridge, Lasso, dan masing-masing bisa punya alpha (misalnya Lasso_a1 = Lasso dengan alpha=1).
<b>train_r2</b>	Nilai R <sup>2</sup> di data training	Mengukur seberapa baik model menjelaskan data latih. Semakin mendekati 1 semakin bagus.
<b>test_r2</b>	Nilai R <sup>2</sup> di data testing	Mengukur seberapa baik model memprediksi data baru (yang tidak dilatih). Ini indikator utama kualitas model.
<b>train_rmse</b>	Root Mean Squared Error (training)	Semakin kecil nilainya, semakin kecil kesalahan prediksi di data latih.
<b>test_rmse</b>	Root Mean Squared Error (testing)	Semakin kecil, semakin akurat prediksi harga di data baru.
<b>test_mape</b>	Mean Absolute Percentage Error (testing)	Persentase rata-rata kesalahan prediksi. Semakin kecil (misalnya 20–30%), semakin bagus.

```

degree      model  cv_r2_mean  cv_rmse_mean
10         2  Ridge_a10    0.632638   109.408590
12         2  Lasso_a1     0.632612   109.432507
11         2  Lasso_a0.1   0.632361   109.446258
9          2  Ridge_a1     0.632325   109.450053
8          2  Ridge_a0.1   0.632284   109.455749
Best Model: Degree 2, Ridge_a10
  luas_tanah  luas_bangunan  jml_kamar  umur  jarak  predicted_harga
  0           120              80        2      5      3       276.433189
  1           300              250        4     10     12       382.524956
  2            80               60        1      2      5       265.680904
  3           450              350        5     20      2       737.639330
  4           200              150        3      8      8       295.938694

  ci_lower    ci_upper
  0  66.163445  486.702933
  1 172.255211  592.794700
  2  55.411160  475.950649
  3 527.369585  947.909074
  4  85.668950  506.208438

```

---

- Setelah menguji berbagai model, hasil cross-validation menunjukkan bahwa model terbaik adalah Ridge Regression dengan derajat 2 dan alpha 10. Model ini memiliki akurasi sekitar 63%.
- Saya kemudian menggunakan model tersebut untuk memprediksi harga beberapa rumah baru.
- Misalnya, rumah dengan luas tanah 120 m<sup>2</sup> dan bangunan 80 m<sup>2</sup> diperkirakan berharga sekitar 276 juta dengan rentang kepercayaan antara 66 juta sampai 486 juta.